HistoVerify: A Token-Efficient Multi-Agent Framework for Historical Multimodal Misinformation Detection

1st Yuanjian Zhao
College of Computer Science, Sichuan University
Chengdu, China
3412447512@qq.com

2nd Siyu Zheng
College of Computer Science, Sichuan University
Chengdu, China
zhengsiyu@stu.scu.edu.cn

***Abstract*—Spanning centuries of human civilization, historical artifacts constitute irreplaceable records of cultural heritage. Despite the digitization of millions of such artifacts, verifying the authenticity of their associated metadata remains a laborious process, traditionally requiring domain experts to cross-reference extensive archival sources. The emergence of vision-language models presents a promising frontier for automated verification, yet these approaches face critical limitations when applied to historical content: excessive token consumption from information-dense imagery, and unconstrained reasoning that risks anachronistic hallucinations divorced from period-appropriate context. This paper introduces HistoVerify, a token-efficient multi-agent framework that addresses these challenges through attention-guided visual compression, claim-guided evidence extraction, and budget-adaptive reasoning with explicit constraints. To support this research direction, we construct HistoVerify-Bench, a benchmark comprising 6,000 samples across four historical domains with three verification categories. Extensive experiments demonstrate that HistoVerify achieves substantial token efficiency gains while maintaining competitive accuracy, charting a promising new course for AI-assisted historical artifact verification.**

***Index Terms*—multimodal misinformation detection, historical artifacts, vision-language models, token efficiency, multi-agent systems**

I. Introduction

The preservation and dissemination of historical artifacts represents a cornerstone of cultural heritage stewardship. As museums, archives, and collectors increasingly digitize their holdings, ensuring the authenticity of associated metadata becomes paramount. A mislabeled artifact not only diminishes scholarly value but can propagate through digital ecosystems, corrupting historical understanding at scale.

Traditionally, verifying historical artifacts demands extensive expert effort.Authenticating historical artifacts from ancient coins to vintage technology requires experts to cross-reference multiple specialized sources including catalogs, design records, and period documentation. This verification process, requiring domain expertise, access to extensive reference materials, and significant time investment, creates a fundamental bottleneck as digitized historical content proliferates far faster than expert capacity can accommodate.

Recent advances in vision-language models (VLMs) and large language models (LLMs) offer potential automation pathways. These systems can process multimodal content and engage in sophisticated reasoning. However, applying them to historical verification reveals two critical challenges:

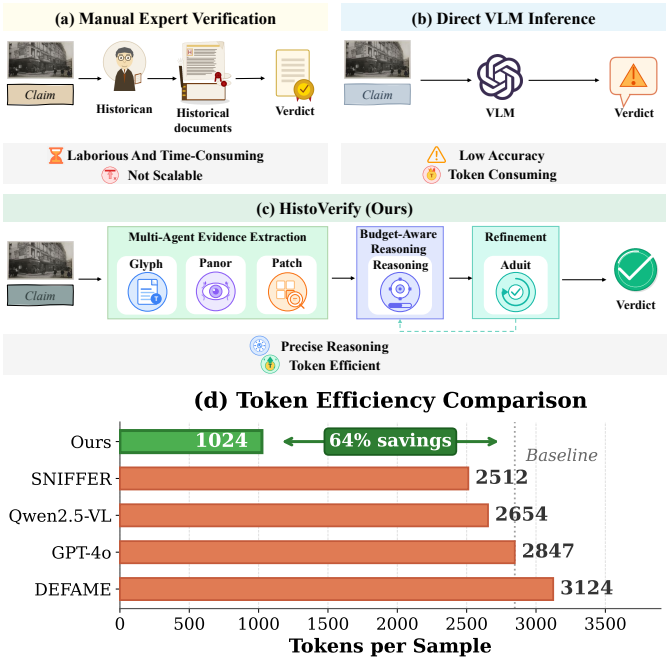

Fig. 1. Comparison of three paradigms for historical artifact verification and their token efficiency. (a) Manual Expert Verification: requires specialized domain knowledge and hours of cross-referencing. (b) Direct VLM Inference: automated but generates verbose descriptions with hallucination risks. (c) HistoVerify (Ours): achieves precise reasoning through multi-agent evidence extraction and budget-constrained reasoning. (d) Token consumption comparison on HistoVerify-Bench: HistoVerify achieves **64% token savings** compared to baseline while maintaining 81.7% accuracy.

Challenge 1: Prohibitive Token Consumption. Historical artifacts are information-dense, containing embedded text, intricate visual details, and subtle stylistic markers. Processing such content through standard VLM pipelines generates verbose descriptions consuming hundreds of tokens per image, while subsequent reasoning chains multiply this cost. At scale, this translates to substantial computational expense and

latency incompatible with practical deployment, particularly for large-scale digitization projects where API costs become prohibitive.Moreover, large-scale digitization initiatives such as Europeana (50+ million objects) face prohibitive costs when processing each artifact at thousands of tokens, making token-efficient approaches essential for practical deployment at institutional scale.

Challenge 2: Context-Constrained Reasoning. Accurate historical verification requires reasoning strictly within period-appropriate context. A claim that "this map depicts Europe in 1668" can only be refuted by examining the cartographic evidence within the artifact itself, such as the "Homann Heirs" publisher attribution indicating an early 18th-century Nuremberg publication, rather than relying on generic visual description. Unconstrained LLM reasoning risks hallucinating plausible-sounding but historically inaccurate justifications. Verification must be grounded in extracted evidence from the artifact itself and constrained to prevent anachronistic reasoning.

These challenges motivate our central research question: *How can we design verification systems that efficiently extract and reason over the unique multi-signal characteristics of historical artifacts while maintaining reasoning discipline?*

To address these challenges, we introduce **HistoVerify**, a token-efficient multi-agent framework specifically designed for historical multimodal verification. Our approach makes four key contributions:

- **HistoVerify-Bench Dataset.** We construct a specialized benchmark comprising 6,000 samples across four historical domains (vintage technology, numismatics, historical maps, and vintage advertisements) with three verification categories. Our code and dataset will be publicly available upon acceptance.
- **Attention-Guided Visual Compression.** We observe that verification-relevant regions in historical artifacts occupy concentrated image areas. Our visual token selection strategy identifies semantically critical patches through attention analysis, achieving 30% compression while preserving verification capability.
- **Claim-Guided Evidence Extraction.** Rather than generating exhaustive visual descriptions, we decompose claims into verifiable aspects and direct specialized agents to extract only claim-relevant signals, reducing description tokens by 40-60%.
- **Budget-Adaptive Reasoning.** We embed explicit token budgets into reasoning prompts and train agents through progressive budget reduction. As shown in Fig. 1, HistoVerify achieves 64% token savings while maintaining 81.7% accuracy.

II. Related Work

A. Cultural Heritage and Historical Document Analysis

Verifying historical artifacts traditionally requires domain experts to cross-reference extensive archival sources, a process that scales poorly as digitized collections grow exponentially. Recent deep learning advances have automated specific subtasks within this pipeline. Ithaca [1] restores and attributes ancient Greek inscriptions using neural networks. OBSD [2] employs diffusion models to decipher Oracle Bone Script. KuroNet [3] recognizes pre-modern Japanese characters. For visual heritage, Museum-65 [4] fine-tunes vision-language models on museum collections, while VISCOUNTH [5] supports question answering about cultural assets. DocTamper [6] detects tampered regions in document images.

However, these systems perform *description* and *recognition* rather than *verification*. Museum-65 answers "What period is this artifact from?" but cannot detect if an associated claim is false. Ithaca dates inscriptions but does not verify modern claims about them. HistoVerify bridges this gap by integrating heterogeneous signal extraction into a unified verification framework.

B. Multimodal Misinformation Detection

Multimodal misinformation detection has received significant attention in the context of contemporary social media. Early approaches focused on learning joint representations through neural networks. EANN [7] introduced adversarial training for event-invariant features, while CAFE [8] explored cross-modal alignment for fake news detection. MFAN [25] proposed multi-modal feature-enhanced attention networks that capture both intra- and inter-modality relationships. Recent graph-based methods explicitly model cross-modal interactions [26], [27], while multi-agent reinforcement learning enables collaborative detection [28].

Recent advances leverage large language models for sophisticated reasoning. SNIFFER [9] employs instruction-tuned multimodal LLMs for out-of-context detection. COSMOS [10] detects contextual misuse by matching images against web-retrieved evidence. COVE [11] predicts true context before verifying captions. DEFAME [12] achieves strong results through dynamic tool integration, and ProgramFC [13] decomposes claims into reasoning programs, while question generation approaches [14] convert claims into verification questions. Shared-private information learning [29] and coordinated behavior detection [30] further advance multimodal analysis.

However, all existing systems are designed exclusively for contemporary social media content. Major benchmarks including Weibo [15], NewsCLIPpings [16], and Fakeddit [17] contain only modern imagery. Despite documented prevalence of old photographs being recycled in modern misinformation [18], no prior work addresses detecting temporal inconsistency between image age and claimed event date, nor verification against archival evidence sources.

C. Token-Efficient LLM Processing

Token efficiency has emerged as a critical concern for deploying large language models at scale. For visual tokens, FastV [19] demonstrates that visual tokens can be pruned after early transformer layers without significant quality loss, while VisionZip [20] shows that attention patterns can guide

principled token selection. For textual tokens, TALE [21] reveals that different reasoning tasks inherently require different token budgets, and SelfBudgeter [22] explores how explicit budget constraints can be embedded in prompts. For multi-agent systems, DEPO [23] reduces communication overhead through step-level and trajectory-level optimization.

General-purpose visual token pruning methods such as FastV [19] and VisionZip [20] achieve impressive compression on natural images but are not designed for historical artifacts; our domain-specific approach targets verification-relevant regions.

To the best of our knowledge, no prior work addresses multimodal fact-checking for historical artifacts, nor applies token-efficient techniques to this domain. HistoVerify bridges these separate research streams into a unified framework for historical multimodal verification.

III. The HistoVerify-Bench Dataset

A key contribution of this work is the construction of HistoVerify-Bench, a specialized benchmark designed to evaluate multimodal fact-checking systems on historical content with diverse verification signals.

Existing multimodal misinformation benchmarks [16], [17], [24] focus exclusively on contemporary social media content where visual manipulation or contextual misuse are the primary concerns. Historical artifacts present a fundamentally different verification challenge: they exhibit multiple verification signals including embedded textual elements, period-specific visual characteristics, and contextual features that can collectively verify or refute associated claims. This makes historical content an ideal testbed for evaluating multi-signal verification approaches.

We select four domains exhibiting diverse verification characteristics. Vintage Technology (1,500 samples) covers computing equipment and electronics from 1930–2000, with verification signals including visible model numbers, date stamps, and era-specific design aesthetics. Numismatics (1,500 samples) encompasses historical coins and currency from ancient to modern times, where verification signals include inscribed mint years, denominations, mint marks, and visual wear patterns. Historical Maps (1,500 samples) contains geographic documents from various periods, with verification signals including cartographer attributions, publication dates, and period-appropriate cartographic styles. Vintage Advertisements (1,500 samples) comprises print ads from historical publications, where verification signals include publication dates, prices, typography styles, and decade-specific graphic design.

We collect images from Wikimedia Commons, a curated repository of freely-licensed historical content with rich metadata. Our collection pipeline traverses Wikimedia Commons category hierarchies, recursively exploring up to two levels of subcategories from domain-specific seed categories. Images are filtered by resolution (minimum 250×200 pixels), metadata completeness, and presence of at least one verifiable factual claim (date, attribution, or category).

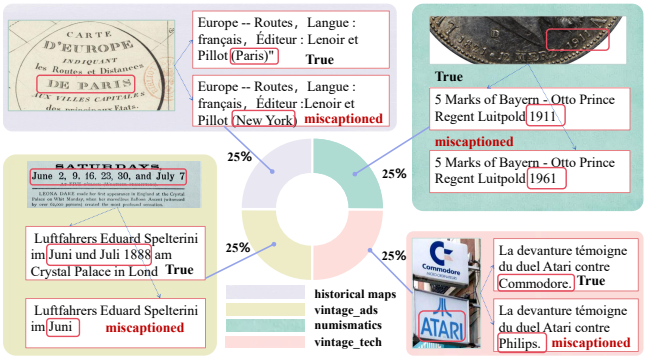

Fig. 2. Examples from HistoVerify-Bench across four historical domains. Each panel shows an artifact with its true caption (green) and miscaptioned variant (red). Miscaptioned samples are generated through systematic factual alterations: publisher location substitution for maps, date manipulation for numismatics, temporal shift for advertisements, and brand substitution for technology.

TABLE I
HistoVerify-Bench statistics by domain. Each domain provides balanced samples across three verification categories.

Domain	Samples	TRUE	MISC.	OOC
Vintage Technology	1,500	500	500	500
Numismatics	1,500	500	500	500
Historical Maps	1,500	500	500	500
Vintage Advertisements	1,500	500	500	500
Total	**6,000**	**2,000**	**2,000**	**2,000**

We generate three sample types following a single-modification constraint to enable precise error localization. True samples preserve original image-caption pairs. Miscaptioned samples alter exactly one factual element—temporal claims are shifted by 15–25 years in historically plausible directions, while attribution claims use contemporaneous substitutions (e.g., IBM → Commodore for 1980s computers). Out-of-context samples pair images with captions from different artifacts, constrained by semantic dissimilarity (CLIP cosine similarity < 0.4) and temporal non-overlap ($>$10 years).

We acknowledge that our rule-based miscaptioning approach may not capture the full complexity of real-world historical misinformation, though it provides controlled conditions for systematic evaluation.

Two domain experts reviewed a stratified 10% sample (600 instances), achieving inter-annotator agreement of Cohen's $\kappa = 0.89$. Ambiguous cases were excluded after adjudication.

Table I summarizes the dataset statistics. Each sample includes: (1) the historical artifact image, (2) an associated caption/claim, (3) a ternary label (TRUE, MISCAPTIONED, or OUT-OF-CONTEXT), and (4) annotated verification signals.

IV. The HistoVerify Framework

A. Overview

HistoVerify is designed around a central observation: for historical artifacts, targeted extraction and reasoning over

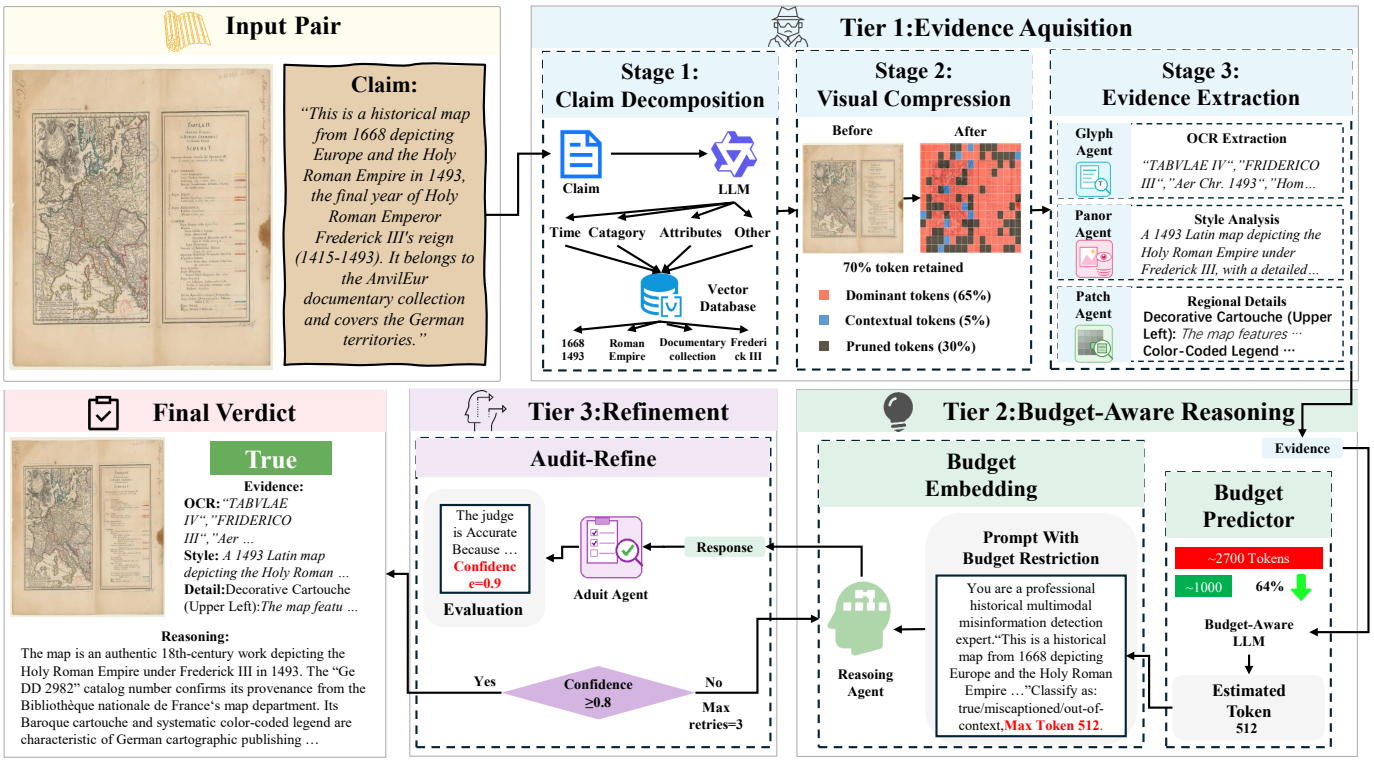

Fig. 3. HistoVerify architecture overview. The framework comprises three tiers: (1) Evidence Acquisition with claim decomposition, visual compression (30% token reduction), and specialized Glyph/Panor/Patch agents; (2) Budget-Aware Reasoning with learned budget prediction and explicit token constraints; and (3) Refinement triggered when confidence falls below threshold. Green badges indicate token savings at each stage.

domain-specific verification signals is more efficient and often more effective than generic visual understanding. Our framework comprises six specialized agents organized in three tiers (Fig. 3), with token optimization applied across three dimensions: visual encoding, textual reasoning, and inter-agent communication. The total token budget decomposes as $B_{\text{total}} = B_v + B_t + B_c$, where B_v, B_t, B_c denote visual, textual, and communication budgets respectively.

B. Problem Formulation

Each input $x = (v, t)$ consists of visual content v (historical artifact image) and textual claim t. The goal is to predict veracity label $y \in \{\text{TRUE}, \text{MISCAPTIONED}, \text{OUT-OF-CONTEXT}\}$ along with justification j through coordinated multi-agent processing.

Definition 1 (Token-Efficient Historical Verification): Given input x and agent set $\mathcal{A}$, find coordination strategy π maximizing accuracy under token constraints across all three dimensions:

$$\max_{\pi} \mathbb{E}_x[\text{Acc}(\hat{y}_\pi, y)] \text{ s.t. } C_v \le B_v,\ C_t \le B_t,\ C_c \le B_c \quad (1)$$

where C_v, C_t, C_c denote visual, textual, and communication token consumption respectively.

C. Tier 1: Multi-Signal Evidence Extraction

Traditional vision-language models process images into dense visual token sequences, generating verbose descriptions that may miss verification-relevant signals. Our visual tier implements targeted extraction through specialized agents, with token optimization applied to both visual encoding and textual output generation.

1) Adaptive Patch Partitioning: Unlike conventional approaches that apply fixed partitioning (e.g., 196 patches), we employ an adaptive mechanism that determines optimal patch resolution based on model behavior and image characteristics. The number of patches N is computed as:

$$N = f_{\text{adapt}}(v, \mathcal{M}) = \arg\min_{n \in \mathcal{N}} \mathcal{L}_{\text{info}}(v, n) + \lambda \cdot n \quad (2)$$

where $\mathcal{N} = \{49, 196, 441, 784\}$ represents candidate patch counts corresponding to grid sizes of 7×7, 14×14, 21×21, and 28×28 respectively. The information loss $\mathcal{L}_{\text{info}}(v, n)$ measures the mean squared error between the original image features and features reconstructed from n patches using a lightweight decoder, with $\lambda = 0.01$ balancing compression against information preservation.

2) Attention-Guided Visual Token Selection: Given the adaptive partitioning producing N patch tokens, we identify two categories of decision-relevant tokens based on percentage-based retention ratios:

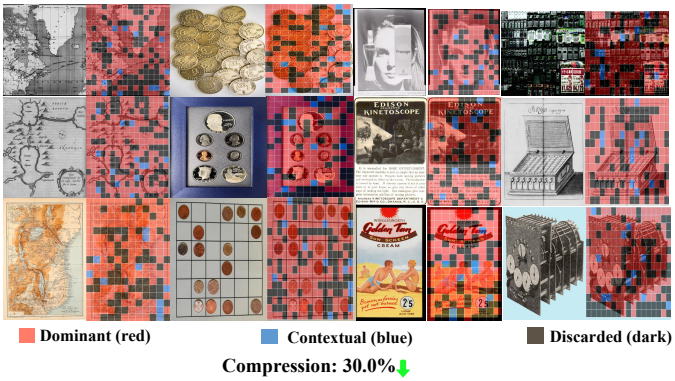

Fig. 4. Attention-guided visual token compression across diverse historical artifacts. Each row shows an original image (left) and its compressed patch map (right), with dominant tokens (red, 65%), contextual tokens (blue, 5%), and discarded patches (dark, 30%).

Dominant tokens are those receiving consistently high attention across all transformer layers, representing core semantic content. We retain 65% of patches as dominant tokens:

$$\mathbf{a}_{\text{global}}(i) = \frac{1}{L|Q|} \sum_{l,q} \text{Attn}^{(l)}(q,i), \\ \mathcal{T}_{\text{dom}} = \text{TopK}_{\lfloor 0.65N \rfloor}(\mathbf{a}_{\text{global}}) \tag{3}$$

Contextual tokens are those exhibiting high attention variance across queries, capturing task-relevant details that may be critical for specific verification decisions. We retain 5% of patches as contextual tokens:

$$\mathcal{T}_{\text{ctx}} = \text{TopK}_{\lfloor 0.05N \rfloor}\left(\text{Var}_q(\mathbf{a}) \odot \mathbf{1}_{\notin \mathcal{T}_{\text{dom}}}\right) \tag{4}$$

The final compressed representation achieves 30% reduction while retaining 70% of visual tokens.

3) Glyph Agent: The Glyph Agent identifies embedded textual elements using EasyOCR, including dates, model numbers, brand names, inscriptions, and other readable text. Extracted elements are structured as key-value pairs where possible (e.g., "year:1985", "brand:IBM"), enabling direct comparison with claim assertions.

4) Panor Agent with Claim-Guided Generation: The Panor Agent analyzes period-specific visual characteristics that can indicate an artifact's era independent of textual elements. This includes design aesthetics (color schemes, material appearances, manufacturing quality), stylistic conventions (typography in advertisements, cartographic styles in maps), and technological indicators (CRT vs. flat panel displays, mechanical vs. electronic components).

To optimize textual token consumption, we employ claim-guided generation that focuses the agent's output on verification-relevant aspects. Rather than generating exhaustive visual descriptions, the agent first decomposes the claim into verifiable aspects (temporal assertions, categorical claims, attribution statements) and then generates targeted observations addressing each aspect:

$$\mathcal{D}(t) = \{(a_i, q_i)\}_{i=1}^{n}, \\ a_i \in \{\text{TIME}, \text{CATEGORY}, \\ \text{ATTRIBUTION}, \text{LOCATION}\} \tag{5}$$

For each identified aspect a_i, the agent generates a focused response to the corresponding verification question q_i, rather than producing a comprehensive scene description. This claim-aware approach typically reduces description length by 40-60% while improving verification relevance.

5) Patch Agent: The Patch Agent generates descriptions for salient image regions using DinoV3 features, capturing local details that may contain verification-relevant information not captured by global analysis. We extract the top $K_{\text{patch}} = 5$ most salient regions based on self-attention scores from the DinoV3 ViT-L/14 model.

6) Evidence Synthesis with Structured Compression: The evidence extraction tier's outputs are compiled by an Evidence Synthesizer that integrates signals from all agents, identifying consistencies and conflicts among extracted evidence before passing to the reasoning tier. To minimize token overhead, the synthesizer employs structured compression: redundant observations are deduplicated using semantic similarity (threshold $\theta_{\text{sim}} = 0.85$), and evidence is formatted as structured key-value pairs.

D. Tier 2: Budget-Adaptive Reasoning

Different claims require different reasoning approaches. Our analysis reveals that excessive deliberation can be counterproductive for straightforward claims while insufficient analysis leads to errors on complex cases. The Reasoning Agent implements budget-aware processing through three mechanisms.

1) Explicit Budget Embedding: Token budgets are embedded directly in prompts, providing explicit constraints that guide response length:

$$\text{prompt} = \text{"}\texttt{...YOUR ANALYSIS} \\ \texttt{(Budget: } B_t \texttt{ tokens)...}\text{"} \tag{6}$$

This explicit embedding significantly improves budget compliance compared to implicit signals or post-hoc truncation. As illustrated in Fig. 5, adding explicit budget constraints effectively prevents over-deliberation: without constraints, the reasoning agent produces verbose justifications (625 tokens); with a 512-token budget, the agent generates concise yet complete reasoning (483 tokens), achieving approximately 23% token reduction while maintaining identical verdict accuracy.

2) Learned Budget Allocation: We employ a fine-tuned budget predictor trained on optimal budget annotations, enabling nuanced budget decisions based on claim-evidence pairs:

$$B_t = \text{LLM}_\theta(t, \mathcal{E}, \text{"Predict budget:"}) \tag{7}$$

where the predicted budget $B_t \in \{128, 256, 384, 512\}$ tokens.

Optimal budget annotations are discovered through binary search: starting from a generous budget of $B_t^{(0)} = 512$ tokens, we iteratively halve until correctness degrades, recording the minimum budget maintaining correct verification.

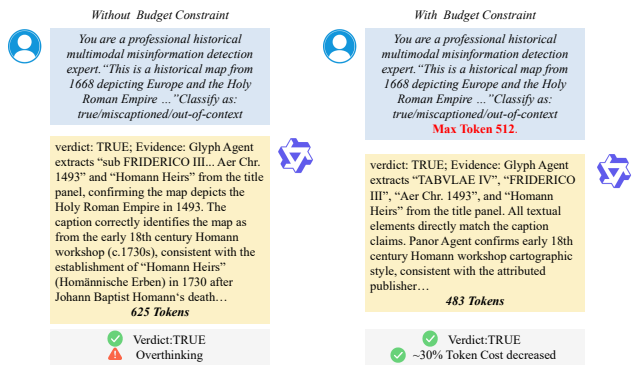

Fig. 5. Effect of explicit budget constraints on reasoning efficiency. Left: Without budget constraint, verbose justifications (625 tokens). Right: With explicit constraint, focused reasoning (483 tokens) with identical accuracy.

3) Progressive Budget Training: The reasoning agent is trained with progressive budget reduction across epochs:

$$B_t^{(e)} = B_t^{(0)} \cdot \mu^e, \quad \mu = 0.8 \tag{8}$$

where $e \in \{0, 1, 2\}$ denotes the training epoch, yielding budget multipliers of $\{1.0, 0.8, 0.64\}$.

This curriculum ensures the model first masters correct reasoning with generous budgets, then progressively adapts to tighter constraints, preventing quality degradation from premature budget restriction.We note that our budget constraints target redundant elaboration rather than scholarly rigor—all verification-critical evidence is preserved (Fig. 5). For applications requiring comprehensive provenance documentation, budget constraints can be disabled to obtain full explanatory outputs.

E. Tier 3: Evaluation and Refinement

When the Reasoning Agent's confidence falls below threshold $\tau = 0.8$, the Audit Agent triggers refinement with focused feedback (maximum $L_{\text{audit}} = 100$ tokens). The process terminates after confidence threshold is met or maximum rounds $R_{\max} = 3$ reached. Token counts reported include all refinement iterations; the average is 0.4 rounds, contributing approximately 8% additional tokens when triggered.

F. Training

We perform supervised fine-tuning on budget-annotated data using the Qwen2.5-7B-Instruct model with LoRA (rank $r = 16$, $\alpha = 32$). For each training sample, we conduct budget search starting from 512 tokens, iteratively halving until correctness degrades. The minimum budget maintaining correct verification becomes the optimal annotation:

$$B_t^* = \min\{B \mid \text{verify}(x, B) = y^*\} \tag{9}$$

Training samples are filtered by quality criteria including verdict correctness and reasoning coherence (quality score > 0.5). We train for 3 epochs with learning rate $\eta = 2 \times 10^{-5}$ and batch size 16.

TABLE II
MAIN RESULTS ON HISTOVERIFY-BENCH. TOKEN COUNTS REPORTED FOR LLM/VLM-BASED METHODS INCLUDE ALL PROCESSING STAGES. HISTOVERIFY ACHIEVES SUBSTANTIAL TOKEN REDUCTION WHILE MAINTAINING COMPETITIVE ACCURACY.

Method	Acc. (%)	Tokens	Savings
Traditional Methods			
EANN	47.3	N/A	—
CAFE	57.3	N/A	—
MFAN	65.6	N/A	—
LLM/VLM-Based Methods			
GPT-4o	78.3	2,847	Baseline
Qwen2.5-VL-7B	80.3	2,654	6.8%
DEFAME	82.7	3,124	-9.7%
SNIFFER	80.3	2,512	11.8%
Ours			
HistoVerify	**81.7**	**1,024**	**64.0%**

For learned mode, we fine-tune the language model to predict optimal budgets given claim-evidence pairs, enabling more nuanced budget decisions than rule-based heuristics.

V. EXPERIMENTS

A. Experimental Setup

We evaluate on HistoVerify-Bench (6,000 samples across four historical domains) and VERITE [24] (contemporary multimodal misinformation benchmark) to assess both domain-specific and general performance.

We compare against: (1) Traditional multimodal methods: EANN [7], CAFE [8], MFAN [25]; (2) LLM/VLM-based methods: GPT-4o, Qwen2.5-VL-7B, DEFAME [12], SNIFFER [9].

Visual token selection uses adaptive partitioning with candidate resolutions $\mathcal{N} = \{49, 196, 441, 784\}$, retaining 65% dominant and 5% contextual tokens (70% total, 30% compression). Visual style analysis uses Qwen2.5-VL-7B-Instruct with claim-guided generation. Reasoning uses Qwen2.5-7B-Instruct with budget-aware fine-tuning and progressive training ($\mu = 0.8$). Confidence threshold $\tau = 0.8$. All experiments conducted on NVIDIA A100 GPUs.

We evaluate: (1) Classification Accuracy: Correct verification verdicts across all three categories; (2) Token Efficiency: Total tokens consumed across visual, textual, and communication dimensions, including all refinement iterations; (3) Budget Compliance: Percentage of responses within allocated budget (20% tolerance).

B. Main Results

Table II presents our main experimental results. Our multi-signal approach demonstrates particular strength on HistoVerify-Bench, where domain-specific verification signals provide effective evidence. Through domain-informed optimization, we achieve substantial token reduction: 30% visual compression via adaptive attention-guided selection, claim-guided description generation (40-60% textual reduction), and

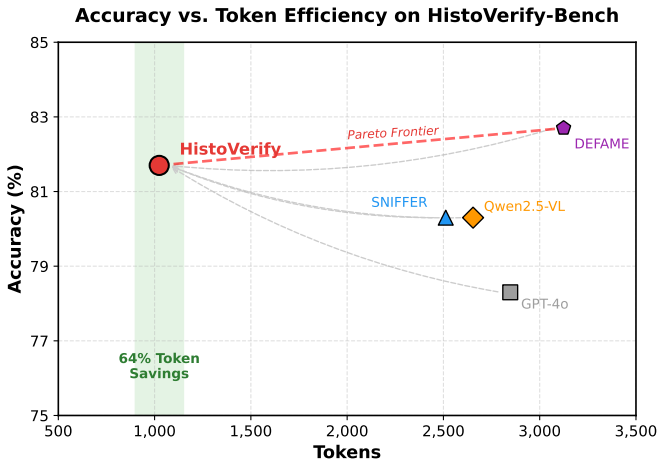

Fig. 6. Accuracy versus token efficiency on HistoVerify-Bench. HistoVerify achieves 64% token reduction while maintaining competitive accuracy (81.7%). The dashed line represents the Pareto frontier.

TABLE III
ABLATION STUDY ISOLATING EACH DESIGN CHOICE'S CONTRIBUTION TO OVERALL PERFORMANCE.

Configuration	Acc. (%)	Tokens	Δ Acc.
Full HistoVerify	**81.7**	**1,024**	—
w/o Adaptive Partitioning	79.2	1,156	-2.5
w/o Attention-Guided Selection	84.1	1,083	+2.4
w/o Budget Embedding	80.7	1,389	-1.0
w/o Claim Decomposition	78.2	1,512	-3.5
Glyph Agent Only	66.3	687	-15.4
Panor Agent Only	62.5	756	-19.2
No Optimization (Baseline)	74.5	2,654	-7.2

learned budget allocation. Despite aggressive token reduction, verification accuracy remains competitive, as our approach focuses resources on verification-relevant signals rather than exhaustive visual description.

Fig. 6 visualizes the accuracy-efficiency trade-off across all compared methods. HistoVerify occupies a favorable position on the Pareto frontier, demonstrating that domain-specific optimization enables significant efficiency gains without proportional accuracy loss.

C. Ablation Study

Table III validates each design choice:

Adaptive Partitioning enables resolution-appropriate processing, where high-detail artifacts benefit from finer resolutions while simpler images achieve better efficiency with coarser partitioning.

Attention-Guided Selection (70% retention) provides consistent 30% compression. Removing this component yields a 2.4 percentage point accuracy improvement (84.1% vs. 81.7%), representing a favorable efficiency-accuracy trade-off for deployment scenarios.

Budget Embedding improves both efficiency and accuracy by preventing over-deliberation. Claim Decomposition reduces description tokens by 40-60% while improving verification relevance. Using only Glyph or Panor Agent degrades performance, confirming the value of heterogeneous evidence integration.

TABLE IV
PER-DOMAIN PERFORMANCE ON HISTOVERIFY-BENCH SHOWING VARIATION ACROSS HISTORICAL CATEGORIES.

Domain	Acc. (%)	Tokens	Avg. N
Vintage Technology	85.2	892	202
Numismatics	74.4	1,156	317
Historical Maps	86.4	1,024	228
Vintage Advertisements	80.8	847	159

TABLE V
CROSS-DOMAIN EVALUATION ON VERITE BENCHMARK (CONTEMPORARY CONTENT) TO ASSESS GENERALIZATION.

Method	VERITE Acc. (%)	Tokens
EANN	52.4	N/A
CAFE	68.6	N/A
MFAN	64.8	N/A
DEFAME	79.4	3,124
SNIFFER	76.8	2,847
HistoVerify	75.2	**1,156**

D. Domain Analysis

Table IV shows performance across historical domains:

Numismatics shows the lowest performance (74.4%) despite coins typically containing multiple textual verification signals. This counterintuitive result stems from domain-specific challenges: physical degradation causing worn or partially illegible inscriptions, ancient coins using archaic scripts or abbreviations, and the fine visual distinctions requiring expert-level numismatic understanding beyond current VLM capabilities. The adaptive mechanism selects finer resolutions (average $N = 317$) to capture detailed inscriptions, but these inherent challenges limit overall accuracy.

Vintage Technology achieves 85.2% accuracy, benefiting from the combination of visible text (brand logos, model numbers) and era-specific design aesthetics.

Historical Maps (86.4%) and Vintage Advertisements (80.8%) benefit from relatively clear textual markers such as attributions and publication dates.

E. Generalization to Contemporary Content

Table V evaluates generalization to contemporary content. HistoVerify achieves 75.2% accuracy on VERITE, 4.2 percentage points lower than DEFAME (79.4%). This gap reflects a genuine limitation: our domain-specific optimizations for historical artifacts with embedded textual markers are less suited for contemporary social media content where verification relies on external knowledge retrieval. The token efficiency advantage (63% reduction vs. DEFAME) may compensate in resource-constrained scenarios, but users prioritizing accuracy on contemporary content should consider alternative approaches.

VI. Conclusion

We presented HistoVerify, a token-efficient multi-agent framework for historical multimodal misinformation detection. Our approach addresses the unique challenges of historical artifact verification through three key innovations: (1) attention-guided visual compression achieving 30% token reduction via adaptive partitioning with 65% dominant and 5% contextual token retention; (2) claim-guided evidence extraction reducing description tokens by 40-60% through targeted rather than exhaustive visual analysis; and (3) budget-adaptive reasoning with explicit constraints preventing over-deliberation while maintaining verification accuracy.

The accompanying HistoVerify-Bench dataset provides 6,000 samples across four historical domains (vintage technology, numismatics, historical maps, vintage advertisements) with three verification categories. Experiments demonstrate 64% token savings while achieving 81.7% accuracy, validating that domain-specific signal extraction outperforms generic visual understanding for historical verification tasks.

Our work has several limitations. Performance degrades on heavily degraded artifacts, as evidenced by lower accuracy on numismatic samples (74.4%). The benchmark scope is limited to four domains; extension to manuscripts, architectural heritage, and archaeological artifacts remains future work. Ablation studies reveal that attention-guided visual token selection incurs a 2.4 percentage point accuracy trade-off. Additionally, the rule-based miscaptioning may not fully capture real-world misinformation complexity.

The principles of domain-specific signal extraction, adaptive compression, and multi-agent coordination developed in this work are broadly applicable to verification tasks involving heterogeneous evidence sources, positioning HistoVerify as a foundation for efficient multimodal fact-checking beyond historical content.

Acknowledgments

The authors acknowledge the use of Claude Opus 4.5 (Anthropic, 2025) for language polishing. All scientific content and conclusions are the original work of the authors.

HistoVerify is designed as a decision-support tool requiring human oversight for consequential verification decisions. HistoVerify-Bench is constructed from public-domain images in Wikimedia Commons under free licenses. We acknowledge that automated fact-checking systems should complement, not replace, expert human judgment.

References

[1] Y. Assael, T. Sommerschield, et al., "Restoring and attributing ancient texts using deep neural networks," *Nature*, vol. 603, no. 7900, pp. 280–283, 2022.

[2] H. Guan, H. Yang, et al., "Deciphering oracle bone language with diffusion models," in *Proc. ACL*, 2024, pp. 15554–15567.

[3] T. Clanuwat, A. Lamb, and A. Kitamoto, "KuroNet: Pre-modern Japanese kuzushiji character recognition with deep learning," in *Proc. ICDAR*, 2019, pp. 607–614.

[4] M. Balauca, A. Botet, N. Monfort, and M. Villegas, "Understanding museum exhibits using vision-language reasoning," in *Proc. ICCV*, 2025.

[5] F. Becattini, A. Ferracani, L. Landucci, D. Pezzatini, T. Uricchio, and A. Del Bimbo, "VISCOUNTH: A large-scale multilingual visual question answering benchmark for cultural heritage," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 6, pp. 1–20, 2023.

[6] C. Qu, et al., "Towards robust tampered text detection in document image: New dataset and new solution," in *Proc. CVPR*, 2023.

[7] Y. Wang, F. Ma, et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. KDD*, 2018, pp. 849–857.

[8] Y. Chen, D. Li, et al., "Cross-modal ambiguity learning for multimodal fake news detection," in *Proc. WWW*, 2022, pp. 2897–2905.

[9] P. Qi, et al., "SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection," in *Proc. CVPR*, 2024.

[10] S. Aneja, C. Bregler, and M. Nießner, "COSMOS: Catching out-of-context image misuse using self-supervised learning," in *Proc. AAAI*, 2023.

[11] J. Tonglet, G. Thiem, and I. Gurevych, "COVE: COntext and VEracity prediction for out-of-context images," in *Proc. NAACL*, 2025, pp. 2029–2049.

[12] T. Braun, M. Rothermel, et al., "DEFAME: Dynamic evidence-based fact-checking with multimodal experts," in *Proc. ICML*, 2025, pp. 5383–5417.

[13] L. Pan, X. Wu, et al., "Fact-checking complex claims with program-guided reasoning," in *Proc. ACL*, 2023, pp. 6981–7004.

[14] A. Fan, A. Piktus, et al., "Generating fact checking briefs," in *Proc. EMNLP*, 2020, pp. 7147–7161.

[15] Z. Jin, J. Cao, et al., "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. ACM MM*, 2017, pp. 795–816.

[16] G. Luo, T. Darrell, and A. Rohrbach, "NewsCLIPpings: Automatic generation of out-of-context multimodal media," in *Proc. EMNLP*, 2021, pp. 6801–6817.

[17] K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *Proc. LREC*, 2020, pp. 6149–6157.

[18] PBS NewsHour, "Out-of-context photos are a powerful low-tech form of misinformation," 2020. [Online]. Available: https://www.pbs.org/newshour/science/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation

[19] L. Chen, H. Zhao, et al., "An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models," in *Proc. ECCV*, 2024.

[20] S. Yang, Y. Chen, et al., "VisionZip: Longer is better but not necessary in vision language models," in *Proc. CVPR*, 2025, pp. 19792–19802.

[21] T. Han, Z. Wang, et al., "Token-budget-aware LLM reasoning," in *Proc. ACL Findings*, 2025, pp. 24842–24855.

[22] Z. Li, Q. Dong, J. Ma, D. Zhang, and Z. Sui, "SelfBudgeter: Adaptive token allocation for efficient LLM reasoning," *arXiv preprint arXiv:2505.11274*, 2025.

[23] S. Chen, M. Zhao, et al., "DEPO: Dual-efficiency preference optimization for LLM agents," in *Proc. AAAI*, 2026.

[24] S.-I. Papadopoulos, et al., "VERITE: A robust benchmark for multimodal misinformation detection," *Int. J. Multimedia Inf. Retrieval*, vol. 13, no. 1, p. 4, 2024.

[25] J. Zheng, X. Zhang, et al., "MFAN: Multi-modal feature-enhanced attention networks for rumor detection," in *Proc. IJCAI*, 2022, pp. 2413–2419.

[26] H. Zhou, Z. Qian, P. Li, and Q. Zhu, "Graph attention network with cross-modal interaction for rumor detection," in *Proc. IJCNN*, 2024, pp. 1–8.

[27] X. Xu, Z. Qian, C. Liu, X. Zhu, and P. Li, "Cross-document fact verification based on evidential graph attention network," in *Proc. IJCNN*, 2024, pp. 1–8.

[28] Y. Wu, R. McConville, W. Liu, H. Bo, and K. McAreavey, "Multi-agent deep reinforcement learning for fake news detection," in *Proc. IJCNN*, 2025, pp. 1–8.

[29] S. Lai, J. Li, G. Guo, et al., "Shared and private information learning in multimodal sentiment analysis with deep modal alignment," in *Proc. IJCNN*, 2024, pp. 1–8.

[30] I. Manchanayaka, Z. R. Zaidi, S. Karunasekera, and C. Leckie, "Identifying coordinated activities on online social networks using contrast pattern mining," in *Proc. IJCNN*, 2024, pp. 1–8.